STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3185–3198
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2171

Assessing uncertainty in reference intervals via tolerance intervals: application to a mixed model describing HIV infection[‡]

Hormuzd A. Katki[*,†], Eric A. Engels and Philip S. Rosenberg

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS 6120 Executive Blvd, Rockville, MD 20852-4910, U.S.A.

SUMMARY

We define the reference interval as the range between the 2.5th and 97.5th percentiles of a random variable. We use reference intervals to compare characteristics of a marker of disease progression between affected populations. We use a tolerance interval to assess uncertainty in the reference interval. Unlike the tolerance interval, the estimated reference interval does not contains the true reference interval with specified confidence (or credibility). The tolerance interval is easy to understand, communicate and visualize. We derive estimates of the reference interval and its tolerance interval for markers defined by features of a linear mixed model. Examples considered are reference intervals for time trends in HIV viral load, and CD4 per cent, in HIV-infected haemophiliac children and homosexual men. We estimate the intervals with likelihood methods and also develop a Bayesian model in which the parameters are estimated via Markov-chain Monte Carlo. The Bayesian formulation naturally overcomes some important limitations of the likelihood model. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: reference interval; tolerance interval; Bayesian statistics; linear mixed models; growth curves

1. INTRODUCTION

A 100α per cent reference interval is the range between the $100(1-\alpha)/2$ and $100(1+\alpha)/2$ percentiles of a random variable in a population; for concreteness this paper sets $\alpha = 0.95$. This interval contains the bulk of the population values and thereby characterizes the typical range of values likely to be observed. Reference intervals are widely used in medicine. For example, a reference interval for blood haemoglobin concentration in newborn infants is 155–232 g/L [1]. Ninety-five per cent of newborns' haemoglobin concentrations are expected to fall within

[*]Correspondence to: Hormuzd A. Katki, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS 6120 Executive Blvd, Rockville, MD 20852-4910, U.S.A.
[†]E-mail: katkih@mail.nih.gov

Received 22 March 2004
Accepted 29 November 2004

this range. Reference intervals plotted over time define a growth curve. Paediatricians use growth curves for height and weight to 'track' young children's development [2]. There is a large literature on reference intervals, reviewed in Reference [3].

Uncertainty in a reference interval estimate can be summarized by an interval that completely covers the true reference interval with 95 per cent confidence (or Bayesian credibility). This interval is a type of tolerance interval [4]. For newborn haemoglobin concentrations, we compute an approximate frequentist tolerance interval of 149–240 g/L, using methods described in Section 4.2. Thus, the true reference interval lies completely within this tolerance interval with 95 per cent confidence.

Reference intervals are used clinically to identify atypical values that require further investigation. For clinical usage, the reference interval endpoints must be estimated precisely. Uncertainty in the endpoints can be assessed by separate confidence intervals for the lower and upper limits. If these confidence intervals are 'sufficiently small', the uncertainty is considered ignorable [5]. Thus, clinical application requires that a reference interval be estimated from sufficiently large sample sizes. In these applications, the estimated reference interval may make use of sophisticated transformations to normality, or semi-parametric and non-parametric methods [3]. In these applications, the marker values for which reference intervals are sought are observed directly for each individual.

In this paper, we consider a different situation. We use reference intervals to characterize how a disease process differs between two affected populations. If the reference intervals for the two populations are markedly different, the disease processes may also be different for the two populations. We are interested in reference intervals for inferred markers, rather than directly observed ones, namely the subject-specific intercepts and slopes of a linear mixed model. The error in inferring the markers must be accounted for by the reference and tolerance interval estimates. In addition, in our applications we have small samples, so that accurate assessment of uncertainty in the reference interval estimates is challenging.

In this setting, a two-sample t-test may not accurately gauge the differences between the distributions of the intercepts and slopes. These markers are not directly observed for each subject, but are estimated from limited data. Furthermore, the t-test assesses the differences between means, but we are also interested in differences between the variances in the two populations. Instead, the reference interval provides a comprehensive measure that reflects the mean and variance of the distribution. Reference intervals from two populations may be disjoint, nested, or partly overlap, and these patterns are of interest. Also, subjects in a cohort with viral loads outside the reference interval may warrant special study.

We derive frequentist and Bayesian estimates of the reference interval, and its tolerance interval, that take the error in inferring the markers into account. The Bayesian model uses non-informative priors for the parameters. Although it is difficult to deal with small samples from a frequentist perspective, this is not a technical issue for the Bayesian method. In addition, the Bayesian method remains valid when the maximum likelihood estimates of the parameters fall on the boundary of the parameter space.

2. MOTIVATION: HIV INFECTION IN ADULTS VS CHILDREN

Human immunodeficiency virus (HIV) viral load is a measure of HIV replication and is used clinically to gauge prognosis and monitor therapy. HIV viral load is defined as the $\log_{10}$ of the

number of HIV virus copies per mL of serum or plasma. The time-course of HIV viral load values differs between persons who acquire the infection as adults and infants who acquire the infection perinatally. Among untreated HIV-infected adults, early viral loads (measured 12–36 months after seroconversion) are typically 3.00–4.00 $\log_{10}$ copies/ml [6]. Viral loads tend to increase slowly over time, although adults vary in their rates of change [7]. In contrast to adults, perinatally infected infants have higher early viral loads, and unlike adults, their viral loads gradually decline from these early values [8].

In a previous paper, we studied typical time trends in HIV viral load among children with haemophilia who contracted HIV from HIV-infected blood products [9]. The time-course of viral loads in these children is of interest because they acquired the infection at an age intermediate between populations of perinatally infected infants and adults. In the Multicenter Haemophilia Cohort Study [10], we identified 22 haemophiliacs who became infected between the ages of 0.7 and 5.2 years. Here we contrast the viral loads observed in these children with those from a cohort of 111 homosexual men from the District of Columbia Gay cohort study who acquired infection as adults [11]. The data were collected prior to the availability of effective therapy; thus the observed trends reflect the natural history of the infection.

A widely used approximation postulates that after 2 years, the mean HIV viral loads within a subject usually reach a stable point after which it progresses linearly over time, so that the intercept and slope characterize the course of infection [7]. We represent the intercept as viral load at 2 years post-seroconversion, and the slope as the yearly change in viral load. Any difference in the distributions of intercepts and slopes between populations may reflect differences in the natural history of infection.

3. REFERENCE INTERVALS AND THE LINEAR MIXED MODEL

If the markers are directly observed, the reference interval is readily constructed using the estimated 2.5th and 97.5th quantiles. This can be done non-parametrically with order statistics or by using a parametric model.

But our markers are the intercepts and slopes of a linear mixed model described below, and these parameters are estimated from limited data. Figure 1 shows viral load trends in four selected HIV-infected children. The single black line in each panel of Figure 1 is the subject-specific regression line. Each subject has potentially different intercepts and slopes, so we use the standard linear mixed model [12].

We model subject i's HIV viral load levels at time t as y_{it}, and denote the corresponding measurement time with origin set to 2 years after seroconversion as x_{it}. Then

$$y_{it} = \beta_{0i} + x_{it}\beta_{1i} + \varepsilon_{it} \tag{1}$$

where $\beta_{0i} \sim \mathrm{N}(\beta_0, \sigma_0^2)$ and $\beta_{1i} \sim \mathrm{N}(\beta_1, \sigma_1^2)$, which are independent of the $(\varepsilon_{i1}, \ldots, \varepsilon_{in})$ that are normally distributed with mean zero and variance σ_ε^2 and the covariance between two times s and t is $\sigma_\varepsilon^2 \rho^{d(s,t)}$ for $d(s,t) = |x_{is} - x_{it}|$. The correlation between β_{0i} and β_{1i} is ρ_{01}. The parameter β_{0i} represents the subject-specific HIV viral load at 2 years after seroconversion. The parameter β_{1i} represents the subject-specific rate of change of HIV viral load per year. We want to construct reference intervals for these two parameters.

The variance components $\sigma_0, \sigma_1, \rho_{01}, \sigma_\varepsilon, \rho$ can be estimated with restricted maximum likelihood (REML). The variance component estimates are then used to obtain weighted least

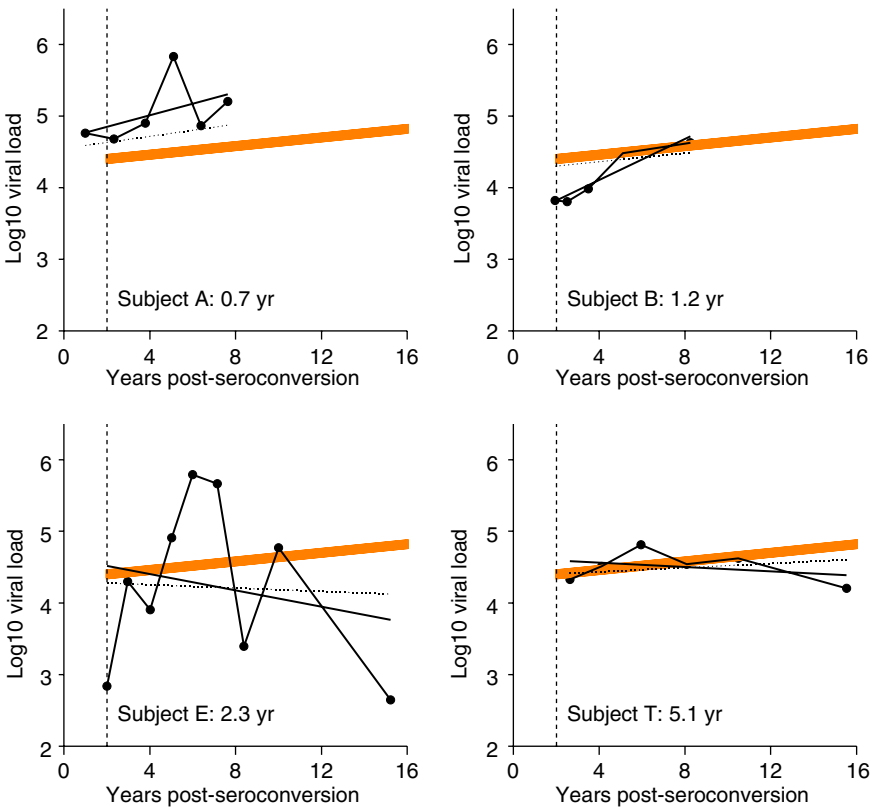

Figure 1. Viral load measurements for four representative HIV-infected children. The thick line is the population average line (the fixed-effects), the thin solid lines are the child-specific regressions, and the dotted lines are the mixed-effects lines.

squares estimates for the fixed effects β_0, β_1. This model can be fit by SAS PROC MIXED [13]. Figure 1 displays the fixed-effect line and mixed-effects lines for four children.

The model-based reference intervals for β_{0i} and β_{1i} are

$$(\beta_j - 1.96\sigma_j, \beta_j + 1.96\sigma_j), \quad j = 0, 1 \tag{2}$$

These intervals are estimated by plugging in estimates of each parameter. These are not confidence intervals; σ_j is the population standard deviation, rather than the standard error of $\hat{\beta}_j$. From here on, when we consider reference intervals in general, we drop the subscripts on β and σ.

4. TOLERANCE INTERVALS FOR REFERENCE INTERVALS

We define the tolerance interval as an interval that completely covers our true reference interval with 95 per cent confidence (or Bayesian credibility). Our tolerance interval is a

random interval $[l, u]$ such that

$$P(F_{\mathrm{L}}(l) \leqslant 0.025 \cap F_{\mathrm{U}}(u) \geqslant 0.975) \geqslant 0.95 \tag{3}$$

where $F_{\mathrm{L}}, F_{\mathrm{U}}$ are the cumulative distribution functions of the lower and upper ends of the reference interval, respectively. This is akin to a (95 per cent, 95 per cent) β-content tolerance interval (see Reference [4, p. 334]), except that our tolerance interval must cover the 2.5th and 97.5th percentiles (i.e. our reference interval, not just any 95 per cent reference interval). The tolerance interval provides inference on the true reference interval akin to those provided by a confidence (or credible) interval for a parameter.

4.1. Other measures of uncertainty for reference intervals

Confidence intervals for the lower and upper limits are often used to assess uncertainty in the reference interval. But these do not make any direct inference on the location and size of the true reference interval. Thus it may be hard to interpret confidence intervals on each endpoint of the interval, especially if they overlap. Another measure of uncertainty is a two-dimensional confidence region for the reference interval endpoints. Although a confidence region makes direct inference on the reference interval, it may be difficult for researchers to easily interpret and communicate general confidence regions. In contrast, the interpretation and communication of tolerance intervals is more straightforward. Merely comparing two reference intervals by testing their endpoints for equality does not yield any information about where the true reference intervals may lie.

Measures of uncertainty are important because the reference interval estimate does not contain the true reference interval with controllable confidence. Thus the reference interval estimate cannot be used to make inference about the true reference interval. Denote the event that the estimated reference interval covers the true reference interval as *Cover*. Then the coverage probability of the estimated reference interval is

$$\begin{aligned} P(\mathrm{Cover}) &= P(\mathrm{Cover}|\hat{\sigma} - \sigma > 0)P(\hat{\sigma} - \sigma > 0) + P(\mathrm{Cover}|\hat{\sigma} - \sigma < 0)P(\hat{\sigma} - \sigma < 0) \\ &= P(\mathrm{Cover}|\hat{\sigma} - \sigma > 0)P(\hat{\sigma} - \sigma > 0) \end{aligned}$$

because $P(\mathrm{Cover}|\hat{\sigma} - \sigma < 0) = 0$. This coverage probability cannot be controlled *a priori*. In large samples, often $\mathrm{median}(\hat{\sigma}) \approx \sigma$ and thus the coverage probability is bounded by 0.5. In the extreme case when σ is known and only $\hat{\beta}$ varies, then the coverage probability is exactly zero. This occurs because if the estimated reference interval covers one of the true reference interval endpoints, then unless $\hat{\beta} = \beta$ exactly, the interval cannot cover the other endpoint. This shortcoming is addressed by the tolerance interval; the tolerance interval is the appropriate widening of the estimated reference interval so as to ensure a controlled probability of containing the true reference interval.

4.2. Approximate frequentist estimation of the tolerance interval

We first consider asymptotic two-sided tolerance intervals of the form

$$\left(\hat{\beta} - 1.96\hat{\sigma} - c\sqrt{\mathrm{Var}(\hat{\beta} - 1.96\hat{\sigma})}, \hat{\beta} + 1.96\hat{\sigma} + c\sqrt{\mathrm{Var}(\hat{\beta} + 1.96\hat{\sigma})}\right) \tag{4}$$

for some critical value c that controls the coverage. We estimate $\mathrm{Var}(\hat{\beta} \pm 1.96\hat{\sigma})$ as

$$\mathrm{Var}(\hat{\beta} \pm 1.96\hat{\sigma}) = \mathrm{Var}(\hat{\beta}) + 3.84\,\mathrm{Var}(\hat{\sigma}) \tag{5}$$

since under REML $\mathrm{Cov}(\hat{\beta}, \hat{\sigma}) \to 0$ [14]. Computationally, SAS PROC MIXED [13] only provides $\mathrm{Var}(\hat{\sigma}^2)$, but by the delta-method [15], $\mathrm{Var}(\hat{\sigma}) \approx (2\hat{\sigma})^{-2}\mathrm{Var}(\hat{\sigma}^2)$.

It is not obvious which choice of c gives the tolerance interval 95 per cent coverage, even under assumptions of asymptotic normality and known variances of $\hat{\beta}, \hat{\sigma}$. For example, if $\mathrm{Var}(\hat{\sigma}) = 0$, then the reference interval reduces to a confidence interval for $\hat{\beta}$, and the usual $c = 1.96$ is the appropriate choice. But if $\mathrm{Var}(\hat{\beta}) = 0$, then the reference interval covers both endpoints or neither, so the reference interval becomes a one-sided confidence interval for $\hat{\sigma}$, and thus $c = 1.645$ is the appropriate choice. To ensure at least 95 per cent asymptotic coverage, we set $c = 1.96$.

The asymptotic tolerance interval has several shortcomings. First, it does not account for finite sample size. Second, it relies on asymptotic normality of $\hat{\beta}, \hat{\sigma}$ which may be a poor approximation since the distribution of $\hat{\sigma}$ is truncated at zero and potentially has a long right tail. Finally, this tolerance interval will be asymptotically invalid if $\hat{\sigma} = 0$. Indeed, if $\hat{\sigma} = 0$, then the reference interval estimate is degenerate and collapses to a point. *A priori*, the true σ is unlikely to be zero, and we would like to incorporate this incontroversial prior information into our modelling. In the next section, we develop a Bayesian approach that overcomes these limitations.

5. BAYESIAN FORMULATION

We start with the mixed model of Section 3 and propose reasonable prior beliefs on the parameters $\beta_0, \beta_1, \sigma_0, \sigma_1, \rho_{01}, \rho$ defined in equation (1). When $\hat{\sigma} = 0$, the Bayesian approach produces non-degenerate reference interval and valid tolerance interval estimates. Furthermore, the Bayesian tolerance interval naturally takes finite-sample uncertainty into account, thus it does not rely on asymptotics nor on the need to specify a critical value c. The Bayesian approach flows from specifying a likelihood function and prior beliefs on the parameters.

5.1. Choice of likelihood

Our likelihood is the standard normal likelihood on the fixed effects given the variance components, combined with REML for the variance components, yielding

$$L(\beta, \sigma, \rho | y) \propto L(\beta | \sigma, \rho, y) \times L(\sigma, \rho | y) \tag{6}$$

where $\beta = [\beta_0, \beta_1]$ and $\sigma = [\sigma_0, \sigma_1, \rho_{01}, \sigma_\varepsilon]$ are vector parameters and y is the vector of observed outcomes. Here

$$L(\beta | \sigma, \rho, y) \propto N(\beta, (X'V^{-1}X)^{-1}) \tag{7}$$

where X is the design matrix of covariates for fixed effects (X' is its transpose). The variance–covariance matrix of y is

$$V = V(\sigma, \rho, y) = XGX' + R$$

where $G = G(\sigma_0, \sigma_1, \rho_{01})$ is the covariance matrix of the random effects, and $R = R(\sigma_\varepsilon, \rho)$ is the covariance matrix of the errors.

We use the standard REML log-likelihood for the variance components [16]. Using the notation of SAS PROC MIXED [13]:

$$l(\sigma, \rho | y) \propto \log |V| + \log |X'V^{-1}X| + (r'V^{-1}r) \tag{8}$$

where

$$r = y - X\hat{\beta}$$

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

and $|\cdot|$ denotes the determinant.

This likelihood is the generally recommended frequentist approach to estimating linear mixed models [17]. An alternative is to use a full normal likelihood, thereby estimating β, σ and ρ jointly. Indeed, Bayesian estimation from the full likelihood is usually easier to compute. However, REML takes account of the implicit degrees of freedom associated with the fixed effects β, while the full likelihood does not [17]. Thus REML often produces more reliable small-sample estimates of the variance components σ and ρ.

5.2. *Choice of priors and estimation*

Since relatively little is known about the time course of HIV viral load in children, we only consider non-informative priors. We consider the priors on each of $\beta_0, \beta_1, \rho, \sigma_0, \sigma_1, \rho_{01}, \sigma_\varepsilon$ to be independent. Although independent priors on intercepts and slopes is a reasonable approximation in our situation, it may not be reasonable in general. However, independent priors on the intercepts and slopes still allows correlated posteriors.

For fixed effects β, we use an improper flat prior. Although this prior is improper, it is considered a reasonable choice because combining data over subjects should be informative about β (see Section 5.4 of Reference [18]) and the posterior for β will be proper. For ρ and ρ_{01}, we use proper non-informative flat priors on the interval $[-1, 1]$ [19]. For the variances σ, we use an improper flat prior. This prior yields proper posterior in this situation [18]. However, a totally flat prior can be unreasonable, and there is no consensus choice of non-informative priors for variances. In Appendix A, we evaluate other priors for σ to assess the sensitivity of the results.

Denoting a chosen prior on σ as $\pi(\sigma)$ and denoting the full posterior as P, the posterior distribution is

$$P(\beta, \sigma, \rho | y) \propto L(\beta | \sigma, \rho, y) L(\sigma, \rho | y) \pi(\sigma) \tag{9}$$

The posterior distribution does not have a closed form, thus Markov-chain Monte Carlo (MCMC) techniques [19] must be used to draw samples $(\beta^{(i)}, \sigma^{(i)}, \rho^{(i)})$ from it. Details of this method are available upon request.

The Bayesian reference interval is estimated by plugging in the posterior mean of β, σ into equation (2). The posterior mean of β, σ is estimated by taking the average of $(\beta^{(i)}, \sigma^{(i)})$. Each draw has a corresponding drawn reference interval of $(\beta^{(i)} - 1.96\sigma^{(i)}, \beta^{(i)} + 1.96\sigma^{(i)})$. The Bayesian tolerance interval takes the form of a highest posterior density interval, so

that it is the smallest interval covering 95 per cent of the drawn reference intervals; see Reference [20] who uses the same procedure but restricts consideration to tolerance intervals symmetric around the posterior means.

6. APPLICATION TO HIV REFERENCE INTERVALS

The model of Section 3 was fit with SAS PROC MIXED [13]. In our data, allowing correlation between intercept and slope does not change the other parameter estimates by much, a likelihood-ratio test for it is insignificant, and the slopes and intercepts estimated from child-specific linear regressions are uncorrelated. Thus we set $\rho_{01} = 0$ in this application.

The results are in Table I. Men and children have similar mean viral load at 2 years after seroconversion, 4.4 $\log_{10}$ copies/ml. The men may have more variation around that mean ($\sigma_0^2 = 0.3$ *versus* 0.1). Children have a higher mean annual increase in viral load ($\beta_1 = 0.03$ *versus* 0.001) but the men may have more variation about their mean ($\sigma_1^2 = 0.004$ *versus* 0.002).

Table II shows reference intervals and their associated tolerance intervals estimated using the REML approach, and the Bayesian approach with flat priors on all parameters. The reference interval for the viral load intercept in children nests inside the corresponding reference interval for the men by nearly one-half $\log_{10}$ on each side. Since one-half $\log_{10}$ is a factor of 3 on the natural scale, the children appear to have much less variation in their initial viral loads than the men. But the reference intervals are estimated with error, and the effect of this error is assessed by the tolerance interval. The children's asymptotic REML tolerance interval is almost identical to that for the men, while the Bayesian tolerance interval for the children is slightly wider than that of the men. Thus, although it appears that children have less variation

Table I. Parameter estimates from usual REML analysis, standard errors in parentheses.

Cohort	β_0	β_1	σ_0^2	σ_1^2	ρ	σ_ε^2
Children viral load	4.4(0.01)	0.03(0.02)	0.1(0.1)	0.002(0.002)	0.3(0.1)	0.6(0.1)
Men viral load	4.4(0.06)	0.001(0.001)	0.3(0.06)	0.004(0.001)	0.3(0.05)	0.4(0.03)
Children CD4 per cent	31(1)	−2(0.3)	0(none)	0.5(0.3)	0.3(0.06)	67(8)

Table II. Reference intervals (RI) and tolerance intervals (TI) for the children and for the men for viral load. Bayesian priors are flat on all parameters.

Cohort	REML RI	Asymptotic TI	Bayesian RI	Bayesian TI
Viral load intercepts				
Children	(3.8, 5.0)	(3.1, 5.7)	(3.6, 5.2)	(2.9, 5.9)
Men	(3.4, 5.5)	(3.2, 5.7)	(3.5, 5.4)	(3.2, 5.7)
Viral load slopes				
Children	(−0.07, 0.12)	(−0.17, 0.23)	(−0.10, 0.17)	(−0.21, 0.30)
Men	(−0.12, 0.12)	(−0.17, 0.17)	(−0.12, 0.13)	(−0.16, 0.20)

Table III. Reference intervals (RI) and tolerance intervals (TI) for the children CD4 per cent. Bayesian priors are flat on all parameters.

Parameter	REML RI	Asymptotic TI	Bayesian RI	Bayesian TI
Children CD4 per cent				
Intercept	(31, 31)	(28, 34)	(25, 38)	(17, 47)
Slope	(−3.1, −0.4)	(−3.9, 0.5)	(−3.8, 0.06)	(−5.4, 1.5)

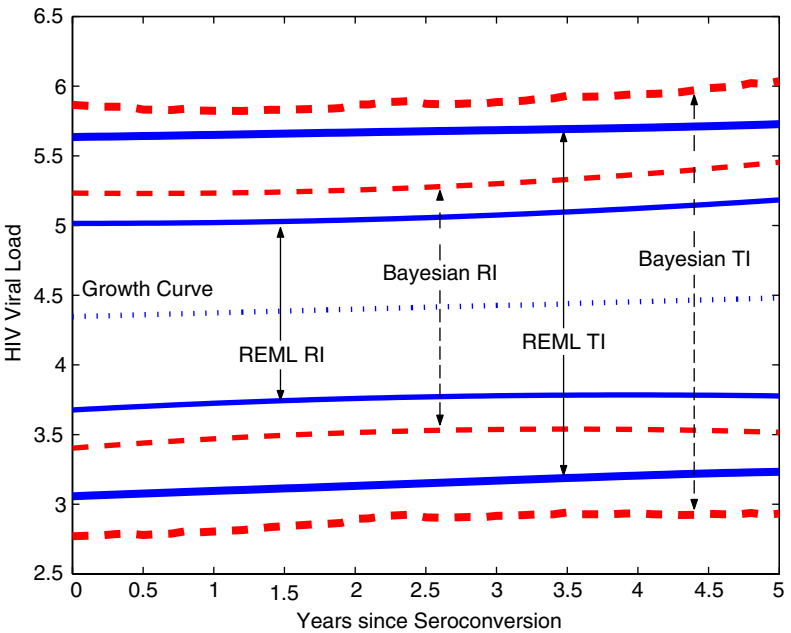

Figure 2. Growth curve, reference intervals (RI) and tolerance intervals (TI) for HIV viral load in haemophiliac children. Bayesian priors are flat on all parameters.

in their initial viral loads than men, there is too much error in the estimates to be sure. Viral load slopes are uncertain in both groups, both for the mean slopes (which do not differ significantly from zero) and for the distribution of subject-specific slopes. For the children's data, the tolerance intervals are much wider than the corresponding reference intervals because our data are limited to 22 children. Since there are 111 men in the adult data, their tolerance intervals are relatively close to their corresponding reference intervals.

We also fit linear mixed models to data on the percentage of lymphocytes that were CD4 positive (CD4 per cent) over time in the same 22 children. CD4 per cent is a marker of immune function. As shown in Table I, the CD4 per cent data produce a zero-estimated variance component for the intercept. As discussed in Section 4, an estimated variance of zero yields invalid reference and tolerance intervals using the REML approach. The Bayesian approach yields reasonable valid intervals. In Table III, the REML reference interval for the intercept degenerates to (31, 31) while the Bayesian estimate is (25, 38).

We plotted reference intervals and tolerance intervals of marker values versus years since seroconversion, yielding a growth curve. Viral load trends in children, as a function of time since seroconversion, are shown in Figure 2. The tolerance intervals are substantially wider than the reference intervals. The Bayesian tolerance intervals appear somewhat irregular due to the discreteness of the MCMC-generated posterior distribution. Details on the computation of growth curve reference intervals and tolerance intervals are presented in Appendix B.

7. DISCUSSION

Reference intervals are useful for comparing characteristics of disease progression between affected populations. We have presented methodology for assessing and conveying uncertainty in reference intervals using tolerance intervals. Other methods of assessing uncertainty, such as a confidence region for the endpoints can be difficult to visualize and communicate. Separate confidence intervals for each endpoint make no inference on the true reference interval. Although the difference between two reference intervals can be assessed by tests of equality on their endpoints, a tolerance interval is needed to determine what other reference intervals could be compatible with the data, or whether a subject with outlying marker values is truly outside reference limits. Our main objectives were to compute reference and tolerance intervals when the markers of interest are inferred with error, to handle a zero-estimated variance component and to account for small-sample variability. Although a frequentist approach can be workable, and a more sophisticated frequentist approach such as a profile likelihood estimate of the upper confidence limit on the variance could work, a Bayesian approach utilizing non-informative priors satisfies all these objectives. Standard MCMC methods quickly yield the Bayesian intervals.

The tolerance interval has been proposed for medical applications where marker values are directly observable [21]. Tolerance intervals for indirectly observable marker values have been considered in the quality control literature. Useful approximate frequentist tolerance interval estimates can be derived [22]. However, these estimates can only be extended to the situation of a zero-estimated variance component by replacing it with its upper bound from a suitably chosen confidence interval. Although advice on choosing the frequentist coverage level has been presented [22], the presence of a zero-estimated variance component remains a source of difficulty. In addition, this upper bound is not computed by SAS PROC MIXED. A Bayesian approach similar in spirit to ours has also been proposed in the quality control literature [20]. Our example is in a biological setting with unbalanced data, correlated errors and includes a slope parameter in the model.

In the viral load example, the intercept reference interval for the children nested inside that of the homosexual men, suggesting that children's viral loads 2 years after seroconversion have less variation. However, the tolerance intervals for both groups were virtually identical; there is too much uncertainty to draw solid conclusions. The tolerance intervals for the men are nearly identical using both approaches, suggesting that both the frequentist and Bayesian tolerance intervals may agree in sufficiently large samples. For the children, the Bayesian tolerance intervals are wider than the asymptotic REML tolerance intervals, because the Bayesian intervals account for the small sample size. In the CD4 per cent example, there is a zero-estimated variance component. The frequentist approach does not produce meaningful reference or tolerance intervals, but the Bayesian approach does.

The standard linear mixed-effects model relies heavily on its distributional assumptions. The normal distribution assumptions must be checked, perhaps via normal *QQ*-plots of the estimated residuals and random effects, and assumptions on the residuals checked via residuals *versus* fitted plots [23]. Assuming homoscedastic errors across subjects requires critical examination [23]. Other important issues are potential informative censoring of viral load measurements by death and potential left-censored viral load measurements [24].

The tolerance interval provides a useful assessment of uncertainty in reference interval estimates, and we believe it should be routinely calculated. The approximate frequentist tolerance interval is easy to compute, but should be used cautiously in small samples, and avoided altogether in case of a zero-estimated variance component. The Bayesian approach appears to be valid in all circumstances. In our application a flat improper prior on the variances is adequate, but we recommend truncating the prior at a reasonable value to exclude impossibly large variances from consideration. In addition, truncation will often stabilize the estimate and improve the convergence of the MCMC. If there is limited prior information, a sensitivity analysis to a variety of non-informative priors is also good practice.

Ultimately, our methods could be used clinically, but much work is required to validate them and convince clinicians of their utility. Astute clinicians consider trends in markers and make clinical judgements. But information gained from formal analysis of real-time patient data might improve patient outcomes. Our methods may apply in other settings where disease markers are measured over time, e.g. screening for prostate cancer with prostate-specific antigen or ovarian cancer with CA125.

APPENDIX A: SENSITIVITY TO PRIORS ON VARIANCES

There is no consensus on non-informative priors for variances. The priors on β_0, β_1, ρ will always be flat as we vary the priors on $\sigma_0, \sigma_1, \sigma_\varepsilon$. All priors are independent and parameterized as in Appendix A of Reference [19]. Since we set $\rho_{01} = 0$ in our application, we do the same here. Table A1 shows the reference intervals and tolerance intervals as estimated with each prior.

The first line is the improper flat prior used in the analysis of Section 6. The second line is truncates the flat prior at a prior agreed upon maximums of $1, 0.5, 2$ for $\sigma_0, \sigma_1, \sigma_\varepsilon$, respectively. This is guaranteed to produce a proper prior and yields the same results as the flat prior. The third line is the Jeffreys prior which is $f(\sigma^2) \propto 1/\sigma^2$. If the data were balanced, this prior would lead to a proper posterior [25]. Due to lack of balance, this prior yields an improper

Table A1. Sensitivity of reference intervals (RI) and tolerance intervals (TI) for the children to differing priors on the variance components.

Prior	Intercept RI	Intercept TI	Slope RI	Slope TI
Flat	(3.6, 5.2)	(2.9, 5.9)	(−0.10, 0.17)	(−0.21, 0.30)
Truncated flat(1, 0.5, 2)	(3.6, 5.2)	(2.9, 5.7)	(−0.10, 0.17)	(−0.22, 0.29)
Jeffreys	(3.9, 4.9)	(3.2, 5.6)	(−0.05, 0.11)	(−0.22, 0.30)
IG(0.5, 0.1)*IG(0.01, 1)*IG(1, 1)	(3.5, 5.3)	(3.0, 5.8)	(−0.09, 0.16)	(−0.17, 0.25)
IG(0.5, 0.1)*IG(0.1, 1)*IG(1, 1)	(3.5, 5.2)	(3.0, 5.7)	(−0.21, 0.30)	(−0.31, 0.42)
Gam(1, 1)*Gam(1, 1)*Gam(1, 1)	(3.6, 5.2)	(2.9, 5.9)	(−0.10, 0.16)	(−0.21, 0.28)
Gam(0.75, 1.33)*Gam(0.75, 0.5)*Gam(1, 1)	(3.6, 5.2)	(2.9, 5.9)	(−0.09, 0.16)	(−0.21, 0.30)

posterior in σ_0 and σ_1 with infinite mass around zero [18]. The resulting intervals are too short. Reassuringly, the improper posterior is easily diagnosed by the MCMC because the chain gets stuck at a point near zero for as many as a thousand iterations.

The fourth and fifth lines are inverse gamma (IG) priors on each of $\sigma_0, \sigma_1, \sigma_\varepsilon$. The only difference between the two lines is the prior of the slope variance σ_1, whose variance is always one but the mean in the fourth line is 0.01 (closer to the REML value) and in the fifth line is 0.1 (far from the REML value). Although these priors are vague, the estimated intervals differ. The problem is that the IG has zero density at zero, thus a 'vague' IG achieves high variance by removing mass near zero and concentrating that mass far from zero. This IG is 'concentrated' yet 'diffuse' and places little mass near zero. Thus the second IG strongly pulls the estimate away from zero and thus cannot be considered non-informative.

The last two lines are two gamma (Gam) priors. Crucially, the Gam is flexible near zero as it can have infinite, finite or zero density at zero. The first choice of Gam is just a exponential distribution with unit mean on all parameters. The second Gam has infinite density at zero (but is proper), huge prior means of $1, 0.375, 1$ for $\sigma_0, \sigma_1, \sigma_\varepsilon$, respectively and has large variance. In spite of the fact that the prior means are far from the REML estimates, both priors lead to estimated intervals that are virtually identical to those from the flat prior. Thus use of the flat prior is justified as it yields intervals that agree with those from a range of reasonable non-informative priors.

We caution that although the Bayesian approach remains feasible when a variance component is estimated to be zero, the likelihood may have little information about the variance component [26]. Thus special care must be taken with the choice of prior, and we strongly recommend a sensitivity analysis akin to this appendix.

APPENDIX B: REFERENCE INTERVALS AND TOLERANCE INTERVALS FOR A GROWTH CURVE

The asymptotic REML and Bayesian reference intervals for each time x_{it} has endpoints

$$\beta_0 + \beta_1 x_{it} \pm 1.96\sqrt{\sigma_0^2 + \sigma_1^2 x_{it}^2} \tag{B1}$$

The REML estimate plugs in its parameter estimates. The Bayesian estimate plugs in the posterior mean for each parameter.

The Bayesian tolerance interval for each x_{it} is easily computed by plugging in each MCMC draw into equation (B1) and then computing the smallest interval completely covering 95 per cent of these intervals [20]. The endpoints of the asymptotic REML tolerance intervals are

$$\hat{\beta}_0 + \hat{\beta}_1 x_{it} \pm 1.96\sqrt{\hat{\sigma}_0^2 + \hat{\sigma}_1^2 x_{it}^2} \pm c\sqrt{\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{it} \pm 1.96\sqrt{\hat{\sigma}_0^2 + \hat{\sigma}_1^2 x_{it}^2}\right)}$$

for some critical value c that we set at 1.96 (see Section 4). We estimate $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{it} \pm 1.96\sqrt{\hat{\sigma}_0^2 + \hat{\sigma}_1^2 x_{it}^2})$ using the delta-method [15] as

$$x^{\mathrm{T}} V_\beta x + 3.84 g^{\mathrm{T}} V_{\sigma^2} g$$

where $x = [1 \ x_{it}]^{\mathrm{T}}, V_\beta$ is the variance–covariance matrix of the $\hat{\beta}_0, \hat{\beta}_1$ estimates, V_{σ^2} is the variance–covariance matrix of the $\hat{\sigma}_0^2, \hat{\sigma}_1^2$ estimates and g is the vector

$$\tfrac{1}{2}(\hat{\sigma}_0^2 + \hat{\sigma}_1^2 x_{it}^2)^{-1/2} \times [1 \ x_{it}^2]^{\mathrm{T}}$$

The matrices V_β, V_{σ^2} are output by SAS PROC MIXED, so this calculation can easily be done. However, if either of $\hat{\sigma}_0, \hat{\sigma}_1$ are zero, then as mentioned in Section 4 this asymptotic frequentist tolerance interval is invalid.

ACKNOWLEDGEMENTS

The authors wish to greatly thank Barry Graubard for his constant encouragement and guidance, and for his comments on earlier versions of this paper. We also wish to thank Robert J. Biggar and James J. Goedert for their valued collaboration in our study of HIV disease progression in children. Finally, we thank the two anonymous referees for their insightful and detailed comments.

REFERENCES

1. Virtanen A, Kairisto V, Irjala K, Rajamaki A, Uusipaikka E. Regression-based reference limits and their reliability: example on haemoglobin during the first year of life. *Clinical Chemistry* 1998; **44**(2):327–335.
2. Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL. 2000 CDC growth charts for the United States: methods and development. *Vital and Health Statistics*, number 246 in 11, National Center For Health Statistics, 2002.
3. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. *Statistical Methods in Medical Research* 1999; **8**:93–112.
4. Wilks SS. *Mathematical Statistics*. Wiley: New York, 1962.
5. Linnet K. Two-stage transformation systems for normalization of reference distributions evaluated. *Clinical Chemistry* 1987; **33**(3):381–386.
6. O'Brien T, Blattner W, Waters D, Eyster E, Hilgartner M, Cohen A, Luban N, Hatzakis A, Aledort L, Rosenberg P, Miley W, Kroner B, Goedert J. Serum HIV-1 RNA levels and time to development of AIDS in the Multicenter Haemophilia Cohort Study. *Journal of the American Medical Association* 1996; **276**:105–110.
7. Hubert J-B, Burgard M, Dussaix E, Tamalet C, Deveau C, Le Chenadec J, Chaix M-L, Marchadier E, Vilde J-L, Delfraissy J-F, Meyer L, Rouzioux C. Natural history of serum HIV-1 RNA levels in 330 patients with a known date of infection. *AIDS* 2000; **14**(2):123–131.
8. Biggar RJ, Janes M, Pilon R, Miotti P, Taha TET, Broadhead R, Mtimivalye L, Kumwenda N, Cassol S. Virus levels in untreated African infants infected with Human immunodeficiency virus type 1. *Journal of Infectious Diseases* 1999; **180**(6):1838–1843.
9. Engels EA, Rosenberg PS, Katki H, Goedert JJ, Biggar RJ. Trends in human immunodeficiency virus type 1 (HIV-1) load among HIV-1-infected children with haemophilia. *Journal of Infectious Diseases* 2001; **184**(3): 364–368.
10. Goedert J, Kessler C, Aledort L, Biggar R, Andes W, White G, Drummond J, Vaidya K, Mann D, Eyster M, Ragni M, Lederman M, Cohen A, Bray G, Rosenberg P, Friedman R, Hilgartner M, Blattner W, Kroner B, Gail M. A perspective-study of human immunodeficiency virus type-1 infection and the development of AIDS in subjects with haemophilia. *The New England Journal of Medicine* 1989; **321**(17):1141–1148.
11. Goedert J, Biggar R, Melbye M, Mann D, Wilson S, Gail M, Grossman R, Digioia R, Sanchez W, Weiss S, Blattner W. Effect of T4 count and cofactors on the incidence of AIDS in homosexual men infected with Human immunodeficiency virus. *Journal of the American Medical Association* 1987; **257**(3):331–334.
12. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
13. SAS Institute Inc. *SAS OnlineDoc*, *Version 8*. SAS Institute Inc.: Cary, NC, 1999.
14. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986; **42**(4):805–820.
15. Lachin JM. *Biostatistical Methods*: *The Assessment of Relative Risks*. Wiley: New York, 2000.
16. Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data (Corr: 94V89 p1572). *Journal of the American Statistical Association* 1998; **83**:1014–1022.
17. Searle SR, Casella G, McCulloch CE. *Variance Components*. Wiley: New York, 1992.
18. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: London, 1995.

19. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall: London, 2000.
20. Wolfinger RD. Tolerance intervals for variance component models using Bayesian simulation. *Journal of Quality Technology* 1998; **30**:18–32.
21. Holst E, Christensen JM. Intervals for the description of the biological level of a trace element in a reference population. *The Statistician* 1992; **41**:233–242.
22. Wang CM, Iyer HK. Tolerance intervals for the distribution of true values in the presence of measurement errors. *Technometrics* 1994; **36**:162–170.
23. Pinheiro JC, Bates DM. *Mixed-effects Models in S and S-PLUS*. Springer: Berlin, 2000.
24. Lyles RH, Lyles CM, Taylor DJ. Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *Applied Statistics* 2000; **49**(4):485–497.
25. Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. Wiley: New York, 1992.
26. Hill BM. Inference about variance components in the one-way model. *Journal of the American Statistical Association* 1965; **60**:806–825.